

Вопросы к экзамену по ММРО

ВМиК, кафедра ММП, осень 2016.

Процедура экзамена

Вы получаете 2 билета и 30 минут на подготовку. Далее вы отвечаете и получаете дополнительные вопросы и задачи. В процессе подготовки к билетам и решения задач вы можете пользоваться заранее заполненным (перед экзаменом) листом А4 с основными моментами из лекций. Лист А4 должен быть написан от руки лично вами. Ничем другим пользоваться нельзя. В процессе ответа экзаменатору, в том числе на его дополнительные вопросы, листом А4 пользоваться нельзя. Для получения положительной оценки вы должны знать ответы на вопросы из теоретического минимума без использования А4.

Комментарии по билетам

Предполагается, что линейная регрессия везде оценивается методом наименьших квадратов. Разложение на смещение и дисперсию (bias variance decomposition) знать пока не нужно, оно будет в следующем семестре.

Билеты

1. Виды обучения: с учителем (supervised), без учителя (unsupervised), частичное (semi-supervised), трансдуктивное. Типы моделей: генеративные, дискриминативные. Типы признаков и типы откликов. Принцип минимизации эмпирического риска. Переобучение, ее зависимость от размера обучающей выборки и сложности модели. Кросс-валидация и A/B тестирование. Дискриминантные функции. Признаковое представление документов - бинарное, TF, TF-IDF.
2. Метод К ближайших соседей. Проклятие размерности. Метрики: L_p , косинусная, Жаккарда. Декоррелирующее преобразование (whitening) и расстояние, которое оно порождает. Взвешенный учет объектов. Пример весов. Отступ и классификация объектов на выбросы, пограничные, типичные, эталонные. Методы фильтрации обучающей выборки для ускорения работы метода.
3. Идея алгоритма оптимизации метода ближайших соседей через структурирование пространства признаков - KD-деревья и ball-деревья. Нахождение расстояния от точки до гиперпрямоугольника и до шара.
4. Задача снижения размерности. Метод главных компонент - 2 определения (через проекции и отклонения), их эквивалентность. Оценка качества аппроксимации отдельной компонентой и первыми К компонентами.
5. Метод главных компонент - определение и итеративный алгоритм их построения. Доказательство, что итеративный алгоритм действительно дает главные компоненты (полученные компоненты удовлетворяют определению).
6. Сингулярное (SVD) разложение - определение. Его связь с главными компонентами. Сокращенное сингулярное разложение порядка К, критерий выбора К. Определить и доказать оптимальность аппроксимации сокращенного сингулярного разложения с точки зрения нормы Фробениуса. Применения сингулярного разложения.
7. Определение решающего дерева. Для решающих деревьев CART: выбор решающего правила в каждом узле для случая классификации/регрессии. Назначение прогнозов узлам дерева в случае регрессии/классификации, симметричных/несимметричных потерь.
8. Определение решающего дерева. Правильные критерии останова наращивания дерева. Обрезка (pruning) для решающих деревьев CART.

9. Оценка классификаторов. Матрица ошибок. Точность, полнота, F-мера. ROC-кривая, мера AUC. ROC кривая для случайного назначения классов. Меры, оценивающие, насколько хорошо классификатор предсказывает вероятности классов.
10. Линейный классификатор. Отступ (margin) и его интуиция. Оптимизационная задача по настройке весов. Основные функции потерь. Метод градиентного спуска и метод стохастического градиента.
11. L_1 и L_2 регуляризация. Какая из них отбирает признаки и почему? Связь минимизации эмпирического риска и принципа максимума правдоподобия. Чему в методе максимального правдоподобия соответствует регуляризация? Вероятностная модель, соответствующая L_1 и L_2 регуляризации.
12. Определение логистической регрессии через вероятности классов. Какой функции потерь она соответствует? Многомерная логистическая регрессия. Функция soft-max.
13. Метод опорных векторов в линейно разделимом и линейно неразделимом случае. Его вывод геометрически. Какой функции потерь и регуляризации он соответствует? Классификация типов объектов в методе опорных векторов.
14. Обобщение методов машинного обучения через ядра. Теорема Мерсера. Операции, не выводящие из класса ядер. Линейное, полиномиальное и RBF ядро - формулы и доказательства, что это действительно ядра. Определение расстояния через ядра.
15. Решение для метода опорных векторов - исходное (с выводом) и его обобщение через ядра.
16. Решение для метода гребневой регрессии - исходное и обобщенное через ядра (с выводом).
17. Матрица штрафов (cost matrix). Байесовское правило минимальной цены. Его упрощение для случая одинаковых штрафов для ошибок на одном классе. Его упрощение для случая единичной матрицы штрафов. Доказать, что последнее упрощение приведет к минимизации числа ошибок классификации.
18. Гауссов классификатор. Сколько параметров имеет этот метод и какие модельные ограничения используются для снижения числа параметров? Доказать, что граница между классами будет квадратичной в общем случае и линейной - при предположении одинаковых внутриклассовых ковариационных матриц.
19. Модель Бернулли и мультиномиальная модель классификации. Являются ли они линейными классификаторами? Сглаживание вероятностей Лапласа. Предположение "наивного Байеса". Как оно работает в случае моделей Бернулли, мультиномиальной и Гауссова классификатора?
20. Ядерное сглаживание для оценки плотности. Виды ядер в одномерном и многомерном случае. Выбор параметра ширины окна (bandwidth). Метод Парзенковского окна. Какому методу оценки плотности соответствует метод K ближайших соседей?
21. Регрессия. Вывод решения для обычной и гребневой регрессии. LASSO и ElasticNet регрессия. Какие регуляризации могут отбирать признаки? Взвешенный учет наблюдений. Алгоритм робастной регрессии. Регрессия Надарая-Ватсона.
22. Сравнение методов классификации: K ближайших соседей, логистическая регрессия с L_1/L_2 регуляризацией, дерево решений, метод опорных векторов с линейным и RBF-ядром, байесовское правило классификации с оценкой плотности - Гауссовой и ядерным сглаживанием (метод Парзенковского окна). Какие методы подвернуты проклятию размерности? Какие методы всегда дают линейные границы между классами? Какие методы используют не все объекты для прогнозирования? Какие методы используют не все признаки для прогнозирования? Насколько методы устойчивы к выбросам? Сравните вычислительную трудоемкость на этапе прогнозирования.
23. Сравнение методов регрессии: K ближайших соседей, регрессия Надарая-Ватсона, линейная регрессия с L_1/L_2 регуляризацией, решающее дерево, робастная регрессия, гребневая (ridge) регрессия с RBF-ядром. Прогноз каких методов зависит линейно от исходных признаков? Какие методы используют не все объекты для прогнозирования? Какие методы используют не все признаки для прогнозирования? Насколько методы устойчивы к выбросам? Сравните вычислительную трудоемкость на этапе прогнозирования.

Теоретический минимум

1. Принцип минимизации эмпирического риска. Типичные функции потерь для регрессии и бинарной классификации.
2. Классификация с помощью дискриминантных функций
3. Определение линейного классификатора (бинарный/многоклассовый)
4. Отступ для классификатора (бинарного/многоклассового)
5. В чем заключается проклятие размерности?
6. Матрица штрафов (cost matrix).
7. Байесовское правило минимальной цены и минимальной ошибки (максимальной апостериорной вероятности)
8. Матрица ошибок. Точность, полнота, F-мера.
9. ROC кривая, AUC.
10. Гауссов классификатор.
11. Метод главных компонент - определение.
12. Обобщение методов через ядра. Типичные ядра: линейное, полиномиальное, Гауссово (RBF).
13. Сингулярное (SVD) разложение - определение.
14. Сокращенное сингулярное разложение порядка K . В каком смысле это разложение дает оптимальную аппроксимацию матрицы?
15. Решающие правила в дереве CART и алгоритм их выбора. Возможные критерии перемешанности классов и откликов в задаче регрессии.
16. Типичные регуляризаторы, какой из может отбирать признаки и почему?
17. Типичные функции потерь для бинарной классификации. Какие из них порождают оптимальные веса, зависящее только от части объектов обучающей выборки?
18. Логистическая регрессия (бинарная, многоклассовая)
19. Метод опорных векторов (бинарный).
20. Ядерное сглаживание для оценки плотности.
21. Гребневая регрессия и вывод оптимальных весов для нее.
22. Алгоритм робастной регрессии.
23. Предположение “наивного Байеса”.